StratusLab

Enhancing Grid Infrastructures with
Virtualization and Cloud Technologies

Review of the Use of Cloud and Virtualization Technologies in Grid Infrastructures

Deliverable D2.1 (V1.2)
11 August 2010

Abstract

This document describes the efforts of the StratusLab project to better understand its target communities, to gauge their experience with cloud technologies, to validate the defined use cases, and to extract relevant requirements from the communities. In parallel, the exercise was used as a dissemination tool to inform people about existing software packages, to increase the awareness of StratusLab, and to expand the our contacts within our target communities. The project created, distributed, and analyzed two surveys to achieve these goals. They validate the defined use cases and provide detailed requirements. One identified, critical issue relates to system administrators' reluctance to allow users to run their own virtual machines on the infrastructure. The project must define the criteria to trust such images and provide sufficient sand-boxing to avoid threats to other machines and services.

StratusLab is co-funded by the
European Community's Seventh
Framework Programme (Capacities)
Grant Agreement INFSO-RI-261552.

Contributors

Name	Partner	Sections
Charles Loomis	CNRS/LAL	All
Mohammed Airaj	CNRS/LAL	All

Document History

Version	Date	Comment
1.0	8 August 2010	Initial version internal review.
1.1	9 August 2010	Minor grammatical and typographical corrections.
1.2	11 August 2010	Corrections based on v1.1 comments. Release for EC.

Contents

List of Figures 6

List of Tables 7

1 Executive Summary 8

2 Introduction 10

2.1 Use Cases 11

2.2 Benefits 12

2.3 Development Roadmap 13

3 Surveys 15

3.1 Methodology 15

3.2 Response Statistics 15

4 Interpretations 18

4.1 Dissemination Aspects 18

4.2 Target Communities 19

4.3 Resource Center Characteristics 19

4.4 Application Characteristics 20

4.5 Cloud Experience and Plans 21

4.6 Virtual Machines and Appliances 22

4.7 Cloud Features 22

4.8 Benchmarks 23

4.9 Additional Use Cases and Encountered Problems 24

5 Conclusions 25

5.1 Extracted Recommendations & Requirements 26

References 30

A System Administrator Survey and Results 32

A.0 Welcome Page 32
A.1 Background Information (Page 1 of 8) 32
A.2 Resource Center Characteristics (Page 2 of 8) 34
A.3 Cloud/Virtualization Plans (Page 3 of 8) 35
A.4 Cloud Experience and Plans (Page 4 of 8) 36
A.5 Virtual Machines & Appliances (Page 5 of 8) 37
A.6 Cloud Features (Page 6 of 8) 38
A.7 Benchmarks (Page 7 of 8) 40
A.8 Use Cases and Encountered Problems (Page 8 of 8) 40

B User Survey and Results 42

B.0 Welcome Page 42
B.1 Background Information (Page 1 of 8) 42
B.2 Application Characteristics (Page 2 of 8) 44
B.3 Cloud/Virtualization Plans (Page 3 of 8) 46
B.4 Cloud Experience and Plans (Page 4 of 8) 46
B.5 Virtual Machines & Appliances (Page 5 of 8) 48
B.6 Cloud Features (Page 6 of 8) 48
B.7 Benchmarks (Page 7 of 8) 49
B.8 Use Cases and Encountered Problems (Page 8 of 8) 50

List of Figures

2.1 The Ecosystem Surrounding the StratusLab Cloud Distribution . 14

List of Tables

2.1 Target Communities 13

3.1 Targets of Survey Announcements 16
3.2 Survey Statistics. 17

4.1 Rankings of Cloud Service Interfaces. 22
4.2 Cloud Feature Category Rankings 23

1 Executive Summary

This document describes the efforts of the StratusLab project to better understand its target communities, to gauge their experience with cloud technologies, to validate the defined use cases, and to extract relevant requirements from the communities. In parallel, the exercise was used as a dissemination tool to inform people about existing software packages, to increase the awareness of StratusLab, and to expand the our contacts within our target communities. The project created, distributed, and analyzed two surveys to achieve these goals.

In terms of dissemination, a large number of people viewed the welcome page of the survey–248 people for the administrator survey and 608 for the user survey. As this page described the goals of the StratusLab project and provided a link to the home page, the awareness of the project has surely increased. In addition, 59 people opted-in to the StratusLab announcement mailing list.

The responses to the surveys indicate that most people within our target communities come from research or educational institutes. Most of the users came from computer science/engineering, bioinformatics, or physics/high-energy physics backgrounds. For the administrators, the vast majority had a connection with high-energy physics; unsurprising as most resources of the European Grid Infrastructure are provided by this community. Geographically, the top three countries for administrators (users) were France, Germany, and Spain (France, Greece, and Switzerland), showing some bias for countries associated with the project.

Of the people completing the survey, a large majority have indicated that they will use or deploy virtualization and cloud technologies, indicating a strong interest in the work of the project. The timescales for use and deployment are broadly compatible with those planned for the project. However, the majority of the users already have some experience with cloud technologies and one-third use clouds regularly. This indicates that the project needs to advance the date at which cloud resources are available to end users.

One concern is that two-thirds of administrators indicated that they will not allow user-generated images to run on their sites. The project must identify the criteria such that site administrators can trust user-generated images. Equally, it must provide an implementation that sufficiently sand-boxes running virtual machines so that they do not pose a threat to other machines or services.

Overall, these surveys validate the scenarios described in the project's Technical Annex and the derived use cases: Grid Services on the Cloud, Customized Environments for Virtual Organizations, Customized Environments for Users, Sharing of Dataset and Machine Images, Provision of Common Appliances, VO/User Services on the Cloud, Deployment of a Group of Machines, and Hybrid Infrastructures. The project has identified five different roles within the target communities and the associated benefits for each. The major limitations that were identified through the surveys related to scalability and sand-boxing. The implementation must scale to the order of 10000 virtual machines. Individual machines must be isolated to ensure that they do not interfere with other machines and services. A list of detailed recommendations and requirements for other activities in the project has been provided.

2 Introduction

The StratusLab partners have, prior to the start of the project, worked closely with users of the European Grid Infrastructure and understand their general requirements. Coupled with the partners' knowledge of virtualization and cloud technologies, these were developed into a set of use cases (or scenarios) that formed the basis of the project's work plan.

The surveys and analysis described in this document continue the partners' engagement with the target communities. They have broadly validated the initial use cases and associated requirements, while helping to define the priorities for the project's roadmap.

The primary aims of the surveys were:

- To understand the target communities of the project, including specifically the system administrators of grid resource centers and scientists interested in using European e-Infrastructures.
- To gauge their knowledge and experience with virtualization and cloud technologies.
- To validate and extend the current set of use cases.
- To enumerate the requirements associated with those use cases.

In addition, these surveys were used to support the broader dissemination efforts of the project:

- To inform members of those communities about existing software and services related to virtualization technologies.
- To increase awareness of the StratusLab project within the target communities.
- To expand the number of direct contacts within the target communities.

This document first reviews the initial use cases and development roadmap defined in the project's Technical Annex. It then describes the methodology used to conduct the surveys and an analysis of the results, which are included in two appendices. The document concludes with a summary of the important points for the development, integration, and deployment activities of the project.

2.1 Use Cases

The primary vision of the StratusLab project is to create a private cloud distribution that permits resource center administrators to deploy grid (and standard) services over the cloud's virtualized resources and that allows scientists to exploit e-infrastructures through a cloud-like interface. This vision emerged from a set of use cases and scenarios developed by the project's partners and based on their experience with existing e-infrastructures. The primary use cases are described below.

Grid Services on the Cloud Grid services are numerous, complex, and often fragile. Deploying these over virtualized resources would allow easy (re)deployment or use of hot spares to minimize downtime. Efficiency of these services is important, so having benchmarks to measure the "virtualization penalty" for real services is equally important.

Customized Environments for Virtual Organizations The computing environment offered by the European Grid Infrastructure is homogeneous regarding the operating system (Scientific Linux is nearly universally used) but inhomogeneous in terms of which software packages are available on a particular site. The first limits the appeal of the infrastructure to those people already using Scientific Linux (or a close relative) and the second increases application failures from missing dependencies. Allowing grid Virtual Organizations (VOs) to develop their own computing environments as Worker Node images would solve both of these issues.

Customized Environments for Users Although allowing VOs to provide virtual machine images increases the utility and reliability of the grid infrastructure, individual users are also likely to require customized environments containing, for example, their own proprietary software and/or data. Extending the ability to create virtual machines to individual users further enhances European e-infrastructures.

Sharing of Dataset and Machine Images Preparing a virtual machine or a dataset image requires significant effort, both in terms of creating it and validating it. Providing a mechanism by which users can share these images avoids duplicated effort and promotes the sharing of knowledge. Solid policy recommendations allowing people to trust those images/dataset must come hand-in-hand with technical solutions to enable the sharing (repositories, access control, security, etc.).

Provision of Common Appliances New virtual machines images are often built from existing ones. Providing simple, stock images of common operating systems lowers the barrier to creating customized images as well as improves the utility of the cloud infrastructure. In addition, appliances for grid services would facilitate their deployment at smaller sites or sites with inexperienced system administrators.

VO/User Services on the Cloud Because of security concerns and lack of tools for controlling network access, VOs and users cannot currently deploy services on the European Grid Infrastructure. VOs often have significant software infrastructures built on top of the grid middleware to provide specialized services to their communities. Users often run calculations that involve workflows or task management. Frameworks to execute those types of calculations usually require a central, network-accessible controller (service) to manage the deployment of tasks and collection of results. Deploying these types of services is possible with a cloud.

Deployment of a Group of Machines Large calculations often involve the deployment of a group of machines. Examples include the deployment of a batch system, Hadoop [2], workflow engines with workers, and BOINC [3]. The cloud should facilitate the deployment of groups of machines to make deployment and configuration of these high-level systems as efficient as possible.

Hybrid Infrastructures All resource centers have a finite amount of computing resources available, although the cloud model aims for the appearance of infinite resources ("elasticity"). To handle peaks and maintain the elasticity of a resource center, the system administrator may want to offer resources from other clouds (e.g. a public cloud like Amazon Web Services [1]) via their cloud. This public/private "hybrid" cloud allows a site to maintain a high-level of service to its users.

2.2 Benefits

The StratusLab cloud distribution will appeal to a wide variety of people including both the users and administrators of resource centers. The project has further divided these two target communities into five different roles, which are described in Table 2.1. People in each of those roles will benefit differently from the use cases described above.

Scientists These people are the least technologically savvy (with respect to computing technology) and simply want to take advantage of a grid or cloud platform as a "black box." They benefit from being able to easily run pre-packaged software and access well-defined datasets. The dynamic nature of the cloud ensures that they can launch their calculations instantly while the elastic nature ensures that they have the resources to finish them quickly.

Software Scientists & Engineers These people have both the scientific and engineering expertise to create software and services that fill scientific needs using the cloud infrastructure. These people benefit from being able to package their services as pre-configured machines, easing use by the scientists but also reducing support effort by avoiding common installation and configuration problems. They may also benefit from wider use of their services by sharing their machine and dataset images through the project's repository.

Community Service Administrators These people bridge the scientific and administrative domains and run general services, such as databases, collaborative tools, or analysis software presented as a service, for their scientific communities. These people benefit from the flexibility of the cloud in terms of service deployment. They can find resources for their services and deploy publicly-accessible services without having to negotiate directly with system administrators. In ad-

Table 2.1: *Target Communities*

User Community	Scientists	End-users that take advantage of existing machine images to run their scientific analyses.
	Software Scientists & Engineers	Scientists and engineers that write and maintain core scientific community software and associated machine images.
	Community Service Administrators	Scientists and engineers that are responsible for running community-specific data management and analysis services.
Resource Center	System Administrators	Engineers or technicians that are responsible for running grid and non-grid services in a particular resource center.
	Hardware Technicians	Technicians that are responsible for maintaining the hardware and infrastructure at a resource center.

dition, they can provide more robust services by providing redundant services in different geographical regions.

System Administrators These people are responsible for running grid and standard services. They benefit from a cloud by being able to migrate services between physical machines (either because of hardware failures or because of load conditions) thus providing more reliable services for their users. They can also more easily deploy hot spares to increase reliability of their services. Prepackaged appliances for grid services will also help inexperienced system administrators reliably deploy these complex services.

Hardware Technicians These technicians maintain the underlying physical hardware. The StratusLab cloud distribution, by separating the running services from the physical infrastructure, will allow these technicians to perform hardware upgrades more quickly by migrating services off of physical hardware needing attention. Similarly, migration can move services from failing hardware, allowing the hardware maintenance to take place during normal working hours.

2.3 Development Roadmap

StratusLab will create a complete, coherent, open-source cloud distribution that allows (grid) resource centers to take advantage of virtualization and cloud technologies. The general idea is that resource centers deploy a private cloud over their physical resources and then deploy their grid and standard services using the virtualized resources of that cloud. Additionally, that cloud will be exposed to users of

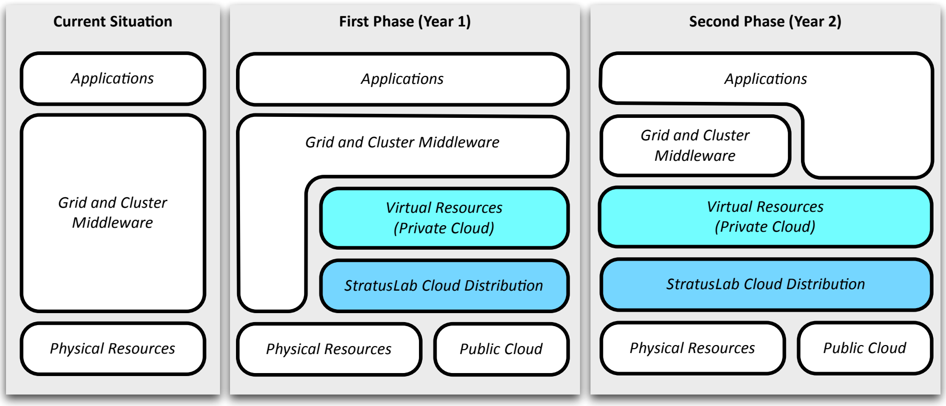

Figure 2.1: *The Ecosystem Surrounding the StratusLab Cloud Distribution*

European e-Infrastructures to give them both cloud and grid services.

Figure 2.1 shows the relationship of the StratusLab cloud distribution to applications, grid middleware, and computing resources as well as its evolution over time. With current grid deployments, the grid middleware resides on physical resources and applications make use of those resources via the grid services.

In the first phase, the project will create the StratusLab cloud distribution that allows grid resource centers to virtualize their physical infrastructure and provide a private cloud. In addition, the StratusLab distribution will allow resources from public clouds (like Amazon Web Services, FlexiScale, ElasticHosts, etc.) to be federated with the site's resources, creating a hybrid cloud infrastructure. The grid middleware will primarily use the virtualized resources for deploying services, although in some cases, direct access to physical resources may be needed for performance reasons. Applications continue to access the resource center's resources via the grid services.

In the second phase, a resource center's private cloud will be directly accessible to end-users through a Cloud API. This allows applications to take advantage of the cloud's benefits, such as the ability to deploy application-level services, easily. In parallel, applications continue to use the available grid services. At this point, the grid middleware should be running on the virtualized resources of the cloud nearly exclusively.

Initially, the first and second phases were intended to correspond to the first and second years of the project. This schedule will need to be modified based on the results of these surveys, in particular, by advancing the availability of the Cloud API to end-users.

3 Surveys

The use cases along with the associated requirements and benefits emerged from the StratusLab participants' experience. While the participants believe that those use cases accurately reflect the needs of the European scientific community, they also realize that they need to be validated by the project's target communities. Consequently, two surveys were created–one for system administrators and another for end-users.

3.1 Methodology

To make the surveys accessible to the widest number of people, they were deployed as web-based surveys using the Zoomerang [25] service. The surveys could be accessed and completed with any modern web browser. The surveys were configured to allow only one response from from a given IP address to avoid duplicates.

The announcement of the surveys was sent via email by members of the project to various mailing lists and contacts. Table 3.1 gives the complete list of recipients for each of the target communities; links to both surveys were always sent in the announcements. The announcement was also placed on the StratusLab [21] and SixSq [20] websites. The dissemination activity tweeted the announcement and it was retweeted by SixSq. Finally, an announcement was placed in the 21 June issue of iSGTW [11]. (A fuller article was planned and written, but unfortunately iSGTW was already full and could not run the article.)

All surveys have an inherent bias and these are no different. As the announcement were made largely through personal contacts; consequently, the respondents are closely related to the EGEE grid community and to administrators using Quattor [18]. Nonetheless, these comprise StratusLab's target communities and the feedback received is highly pertinent for guiding the project's development roadmap.

3.2 Response Statistics

The text of the surveys along with the raw data are available in Appendices A and B. Both surveys included a general section followed by specific questions concerning the use or planned use of virtualization and cloud technologies. Those that indicated that they were not interested in those technologies finished the survey at that point. Those are labeled "Partial Response" in Table 3.2. Those labeled "Full Response" also completed the later section. Those labeled "Incomplete" stopped

Table 3.1: *Targets of Survey Announcements*

	Admin. Survey	User Survey
DCI mailing list	X	X
VENUS-C [22]	X	X
OpenNebula Community Group (LinkedIn)	X	X
HPCcloud.org (mailing list, LinkedIn group)	X	X
OpenNebula Users	X	X
EGI-TF Cloud Program Committee	X	X
HEPiX working group	X	
Quattor Mailing List	X	
WLCG [24]	X	X
LCG-FR [13]	X	X
CernVM [5]	X	X
EGI Participants	X	X
BIGGrid	X	X
D-Grid	X	X
BEGrid	X	X
SEE-GRID	X	X
Greek NGI	X	
FranceGrilles	X	
Grid-Ireland	X	
Spanish NGI	X	
Swiss NGI	X	
EGEE NA4 Regional Contacts	X	X
EGEE NA4 Steering Committee		X
ROSCOE Partner Contacts		X
French Bioinformatics Community		X
EGEE Biomed Contacts		X
HealthGrid 2010 [9] Cloud Presenters		X

Table 3.2: *Survey Statistics*

	Admin.	User
Days Open	30	30
Pages	8	8
Questions	41	36
Visits	248	607
Incomplete	12	43
Partial Response	22	56
Full Response	15	43

before completing the first section.

Table 3.2 contains some statistics concerning the two surveys. They were open for one month (30 June to 30 July), although almost all of the complete responses ("Partial" and "Full") came within 1 week of the original announcement–20 of 23 responses for the administrator survey and 51 of 57 for the administrator survey. The last response was 15 July, indicating that the iSGTW announcement was not effective in generating new responses. Probably a better strategy for the future is to leave the survey open for two weeks with a direct reminder after the first week.

The surveys were rather lengthy and took around 15 minutes to complete. This may have discouraged some of the respondents from completing the survey. The large number of incomplete responses compared to the partial and full responses indicates that this is indeed an issue. However, it would be difficult to significantly shorten the surveys while maintaining the usefulness of the results. For both the user and administrator surveys, approximately 10% of the people visiting the survey completed it.

4 Interpretations

This chapter interprets the results of the survey and identifies important requirements that will allow the project to prioritize its work. The chapter first discusses the dissemination aspects of the surveys. It then analyzes each section of the surveys, treating the administrator and user surveys in parallel when there is a significant overlap in the questions.

References to specific questions are indicated like this: Q5. The context identifies to which survey the reference refers. Recommendations or requirements that come from the responses are indicated in italic and collected in the conclusion section.

4.1 Dissemination Aspects

There were three broad goals to support the broader dissemination efforts of the project:

- To inform members of those communities about existing software and services related to virtualization technologies.
- To increase awareness of the StratusLab project within the target communities.
- To expand the number of direct contacts within the target communities.

Results of the surveys confirm that progress has been made along these lines.

As these surveys were web-based, they could include links to other web sites. Questions that mentioned specific technologies systematically included links to relevant web sites. We could not track how many of those links were actually used. Nonetheless, we hope that interested respondents followed them to find out more information about virtualization, cloud, and other related technologies.

The welcome pages for the administrator and user surveys were designed to describe the StratusLab project and goals as well as the surveys themselves. All of those visiting the surveys should at least be aware of the project and have seen a link to the project's web page. The administrator and user surveys were visited 248 and 607 times, respectively. These visits include the number partial and complete survey responses, but do not include multiple visits from the same IP address. (Geographic information is not available from these visits.) Consequently, information about the StratusLab project has reached a significant number of people within our target communities.

As a result of these surveys, 59 people have agreed to be on the StratusLab announcement mailing lists (including both incomplete and complete responses). There were 16 people who responded to the administrator survey that opted-in and 48 people from the user survey. Five people responded to and opted-in on both surveys.

Overall, the project has increased the number of direct contacts interested in the StratusLab results and increased awareness of the project within the targeted communities through these surveys.

4.2 Target Communities

Both administrators (Q4) and users (Q4) were asked to identify the institute at which they work. In both cases, about 89% of the respondents said they worked at a research or educational institute. The remainder were from commercial or governmental entities. *Although the research community is the primary target of the project, the project needs to make a stronger effort in contacting commercial enterprises.*

Both asked people to identify their scientific or commercial domain of activity. For the users (Q7), the three largest categories were Computer Science/Engineering (16 responses), Bioinformatics (15), and Physics/High-Energy Physics (9). For the administrators (Q6), the largest response was Physics/High-Energy Physics (12). This bias towards High-Energy Physics in the administrator responses is probably a reflection of the fact that most sites in EGI are High-Energy Physics laboratories or institutes.

The users (Q6) were asked to classify themselves as an End-User/Scientist, Developer/Engineer, Service Administrator, or Computer Science Researcher. Not surprisingly, most people identified themselves as an End-User/Scientist (23/56), with fewer Developer/Engineers (17/56) and Service Administrators (3/56). There were a significant number of Computer Science Researchers (13/56) that responded to the survey.

4.3 Resource Center Characteristics

Only the administrators were asked questions concerning their resource centers. These questions were intended to determine what types of services and tools need to be taken into account for the StratusLab development.

Q8 asked about the primary operating system used for the resource center's physical hosts. Overwhelmingly, RedHat-based linux systems (19/22) were used. The remainder were Debian-based systems. No other operating systems were identified as being used on physical resources in the resource centers. *The project should support installation of the cloud distribution on RedHat and Debian systems, with RedHat systems having a much higher priority.*

The survey identified Quattor and Puppet [17] as being the two most popular site configuration and management tools used within our target community. *Integration of the cloud distribution with automated site configuration and management tools should be demonstrated with Quattor and/or Puppet, with Quattor being the more popular.*

Storage technologies will play an important role in the cloud distribution. Consequently, Q10 asked about the file systems that have been deployed on a site. As expected, local file systems and the Network File System (NFS) comprise the majority of responses. However, this question also showed that a significant number of sites (1/4 to 1/3) have experience with more advanced file systems like GPFS [10] and Lustre [14].

From Q11 and Q12, one sees that the majority of the sites are currently running the gLite [8] grid middleware and that they deploy the standard services: a computing element, a storage element, a file catalog, and a VOBOX. Other gLite services were also indicated, but at a lower level. *Demonstrations of grid services over the cloud should initially target core services of the gLite middleware.*

4.4 Application Characteristics

Only the users were asked questions concerning their applications. These questions were intended to give the project a good idea of the types of applications that will run on the cloud infrastructure as well as characteristics of those applications that could be used for StratusLab benchmarks.

RedHat-based systems (Q8) were strongly favored by the users (26/56), albeit less strongly than for the administrators. Debian-based systems were the next most popular (13/56). Unlike the administrators, the users identified a significant number of other operating systems. *The cloud distribution must supply stock images for popular Red-Hat and Debian-based systems. The cloud infrastructure must be as operating system neutral (with respect to running virtual machines) as possible to maximize its utility.*

Q9 showed that the most common type of application is based on a sequential executable (36/56). However, multi-threaded executables, shared memory applications, and parallel applications were also common. *The application benchmarks must cover all of these types of applications: sequential, multi-threaded, shared memory, and parallel.*

Users indicated (Q10) that most of their analyses involved multiple jobs or tasks, with manual submission, master/worker frameworks, and workflow frameworks being the most commonly used mechanisms for control. *The application benchmarks should include workflow and master/worker applications.*

Q11–Q13 asked about the data input, output, and time requirements for individual applications. The most popular answers were 100 MB of input, 100 MB of output, and 1 hour of CPU time; however, the range of answers was very large. *The application benchmarks must be parameterized to allow a wide range of input sizes, output sizes, and running times to be evaluated.*

Concerning the privacy of data stored on a cloud (Q14), nearly two-thirds of the respondents do *not* have privacy, contractual, or confidentiality constraints for their data. The remaining third *do* have such constraints. Those with constraints indicated that access control lists and encryption would suffice for these data. *The StratusLab cloud implementation must include access control mechanisms for stored data and must permit the use of encrypted data.*

By far, the most important mechanisms for accessing data from within applications were POSIX file access and remote file I/O. A significant number, however, also indicated that block access to data and access to object/relational databases is important. *The cloud must allow both file and block access to data, although file access is by far more important. The cloud must allow access to data stored in object/relational databases.*

4.5 Cloud Experience and Plans

After the previous sections, both surveys asked if the administrator (user) had deployed (used) or was planning to deploy (use) virtualization and/or cloud technologies. A strong majority of both administrators (68%) and users (77%) indicated that they would. Those responding positively were asked the remaining questions of the survey.

When asked about their timescales for use of virtualization and cloud technologies, 40% of administrators (Q15) and 37% of users (Q17) indicated that they were already using them. For administrators, all of them indicated that they would be deploying the technologies within the next 12 months. For users, the peak seems to be within the next 12-24 months. This corresponds well with the plans for the project. *Short-term work (<12 months) should concentrate on developments for deploying cloud infrastructures and longer-term work should concentrate on their use.*

Users were asked (Q18) what type of cloud(s) they would be using. From the responses it is clear that both public and private clouds will be used extensively within the community. This implies that standardization is important to minimize the disruption in moving between different clouds. Interestingly for StratusLab, the most popular response to this question was "your own resources configured as a cloud". *The StratusLab distribution must be simple enough for users themselves to configure their own resources as a cloud.*

The following question (Q19) indicated that the majority have already used a cloud infrastructure and that more than a third use a cloud infrastructure regularly. The two most popular public clouds (Q20) are Amazon Web Services and Google App Engine.

Administrators were asked about the techniques and technologies they intend to use. Q17–Q18 indicate that nearly all sites will use hardware virtualization and most sites will also use para-virtualization. *The StratusLab distribution must allow both full-virtualization and para-virtualization to be used.* The three most popular hypervisors are KVM [19], Xen [6], and VMWare ESXi [23]. Of the open-source virtualization managers (Q20), OpenNebula [16] and Eucalyptus [7] were the most

Table 4.1: *Rankings of Cloud Service Interfaces*

Admin.	User	
1	2	Command Line Interface
2	4	HTTP(S)-base API
3	1	Programming API
4	3	Browser-based Interface
5	5	Web Service API

used with OpenNebula receiving a higher ranking.

Both administrators (Q22) and users (Q22) rated various interfaces to a cloud infrastructure. The users generally rated these interfaces as being more useful than system administrators. They both agreed that a Web Service interface is the least useful but disagreed on the rankings of the others. Table 4.1 show the rankings of the proposed cloud service interfaces. *The cloud service must have a command line interface and a programmable API.*

Administrators were further asked what types of services they would run within virtual machines. A strong majority will be running both grid services (Q24) and standard site services (Q23) using virtualization. They did indicate (Q25) that they will be monitoring the performance of services and run services on physical machines where necessary. A particular worry was that IO performance would not be sufficient to permit storage services to be run on the cloud. *The cloud distribution must allow a broad range of grid and standard services to be run. Quantitative performance evaluations must be done to understand the penalties in using virtualization.*

4.6 Virtual Machines and Appliances

The vast majority of users (Q23) and administrators (Q26) create their virtual machine images themselves or trust a known colleague within their community to do it. Moreover, users (Q24) and administrators (Q27) most often manually create those images from operating system media or distributions. Both groups were asked for criteria for trusting images produced by someone else, but neither group provided detailed criteria. Two-thirds of the system administrators said that they would not allow user-generated images to run on their sites (Q28). Building trust in user-generated images is critical for making the best use of cloud infrastructures. *The project must determine the criteria by which administrators and users can trust machine images.*

4.7 Cloud Features

Both the user survey (Q26–30) and the administrator survey (Q30–Q34) presented a long list of cloud features to be ranked. (See the appendices for the complete

Table 4.2: *Cloud Feature Category Rankings*

Admin.	User	
1	3	Monitoring Features
2	1	General Features
3	2	Data-related Features
4	4	Virtual Machine Control
5	5	Networking Features

list of features.) These were organized into five categories: general features, data-related features, networking features, virtual machine control, and monitoring features. Generally within each category there was not much variation within the rankings, although between categories there was more of a difference.

Table 4.2 shows the relative importance of the various categories to administrators and users. The only major difference between the rankings from administrators and users is that administrators valued monitoring aspects more highly than general and data-related features. This is understandable as a major part of their work involves understanding how their computing resources are being used.

Additional features requested include scalability (to $>$10000 virtual machines) and sand-boxing of running machines. Administrators were also asked to rank the integration of the cloud management with automated site management tools. They ranked this even higher than the monitoring features.

The following are recommendations coming from these questions:

- *The project should consider all features listed in the surveys as valid requirements.*
- *Integration with site management tools is a critical short-term requirement.*
- *The cloud implementation must scale to O(10000) virtual machines.*
- *The implementation must sufficiently sandbox running machine images to prevent unintended or malicious behavior from affecting other machines/-tasks.*

4.8 Benchmarks

The project has already defined seven application-oriented benchmarks for evaluating the performance of the cloud implementation:

- CPU-intensive: High-CPU requirements but little or no input and output data.
- Simulation: Small input, but significant output data with high-CPU requirements.

- Analysis: Large input data but relatively small output data.
- Filtering: Large input and output data.
- Shared Memory: OpenMP-like programs
- Parallel: MPI-like programs
- Workflow: Multiple interdependent tasks.

The users (Q32) ranked all of these highly with similar scores (3.7–4.1). The administrators (Q37) showed more variation (3.1–3.9). However, it is clear that all of these are valid benchmarks to consider. The administrators also pointed out that standard performance tests such as HEP-SPEC [4] for CPU, Iozone[12] for disk IO, and iperf [15] for network performance should also be run. *The project must create application benchmarks (CPU-Intensive, Simulation, Analysis, Filtering, Shared Memory, Parallel, and Workflow) to measure quantitatively the performance of the cloud implementation for realistic applications. Performance benchmarks should also be created using packages like HEP-SPEC, Iozone, and iperf for CPU, disk IO, and network performance, respectively.*

4.9 Additional Use Cases and Encountered Problems

The last three questions of the both surveys allowed respondents to provide additional use cases, describe problems they have encountered with virtualization/cloud technology, and to provide additional comments. None of the responses were particularly pertinent to the planning of the project's work plan.

5 Conclusions

This document describes the efforts of the StratusLab project to better understand its target communities, to gauge their experience with cloud technologies, to validate the defined use cases, and to extract relevant requirements from the communities. In parallel, the exercise was used as a dissemination tool to inform people about existing software packages, to increase the awareness of StratusLab, and to expand our contacts within our target communities. The project created, distributed, and analyzed two surveys to achieve these goals.

The two surveys targeted the administrators of grid resource centers and the users of European e-infrastructures. They were web-based surveys managed with the Zoomerang service. This proved to be an effective method for wide distribution of the surveys and for later analysis of the results. The surveys were open for 30 days (30 June to 30 July), although most of the respondents completed the survey within the first week. In the future, the survey should only be left open for two weeks and a reminder should be sent after the first week. Additionally, the vacation periods of July and August should be avoided to maximize the number of responses.

In terms of dissemination, a large number of people viewed the welcome page of the survey–248 people for the administrator survey and 608 for the user survey. As this page described the goals of the StratusLab project and provided a link to the home page, the awareness of the project has surely increased. In addition, 59 people opted-in to the StratusLab announcement mailing list. The surveys systematically provided links to referenced software packages and technologies. Although we do not know how many of those links were followed, they no doubt increased general awareness of them somewhat.

The responses to the surveys indicate that most people within our target communities come from research or educational institutes. Very few responses came from commercial enterprises–something the project should try to remedy. Most of the users came from computer science/engineering, bioinformatics, or physics/high-energy physics backgrounds. The last two are unsurprising as the high-energy physics community is the primary user of the European Grid Infrastructure (EGI) and the project incorporates people from the bioinformatics community. The large number of computer science and engineering users was unexpected. For the administrators, the vast majority had a connection with high-energy physics; again unsurprising as most resources of EGI are provided by this community. Geographically, the top three countries for administrators (users) were France, Germany, and Spain (France, Greece, and Switzerland), showing some bias for countries associated with the project.

Of the people completing the survey, a large majority have indicated that they will use or deploy virtualization and cloud technologies, indicating a strong interest in the work of the project. When asked about the timescales for use or deployment, the administrators all said within 12 months and the users generally within 24 months. This validates the project's plan to concentrate on administrator needs in the first year and user needs in the second. However, the survey also indicated that the majority of the users already have some experience with cloud technologies and that one-third use clouds regularly. This indicates that the project needs to advance the date at which cloud resources are available to end users.

One concern is that two-thirds of administrators indicated that they will not allow user-generated images to run on their sites. The project must identify the criteria such that site administrators can trust user-generated images. Equally, it must provide an implementation that sufficiently sand-boxes running virtual machines so that they do not pose a threat to other machines or services.

Overall, these surveys validate the scenarios described in the project's Technical Annex and the derived use cases: Grid Services on the Cloud, Customized Environments for Virtual Organizations, Customized Environments for Users, Sharing of Dataset and Machine Images, Provision of Common Appliances, VO/User Services on the Cloud, Deployment of a Group of Machines, and Hybrid Infrastructures. The project has identified five different roles within the target communities and the associated benefits for each. The major limitations that were identified through the surveys related to scalability and sand-boxing. The implementation must scale to the order of 10000 virtual machines. Individual machines must be isolated to ensure that they to not interfere with other machines and services. Detailed recommendations and requirements are listed in the following section.

5.1 Extracted Recommendations & Requirements

1. Although the research community is the primary target of the project, the project needs to make a stronger effort in contacting commercial enterprises.

2. The project should support installation of the cloud distribution on RedHat and Debian systems, with RedHat systems having a much higher priority.

3. Integration of the cloud distribution with automated site configuration and management tools should be demonstrated with Quattor and/or Puppet, with Quattor being the more popular.

4. Demonstrations of grid services over the cloud should initially target core services of the gLite middleware.

5. The cloud distribution must supply stock images for popular Red-Hat and Debian-based systems.

6. The cloud infrastructure must be as operating system neutral (with respect to running virtual machines) as possible to maximize its utility.

7. The application benchmarks must cover all of these types of applications: sequential, multi-threaded, shared memory, and parallel.

8. The application benchmarks should include workflow and master/worker applications.

9. The application benchmarks must be parameterized to allow a wide range of input sizes, output sizes, and running times to be evaluated.

10. The StratusLab cloud implementation must include access control mechanisms for stored data and must permit the use of encrypted data.

11. The cloud must allow both file and block access to data, although file access is by far more important.

12. The cloud must allow access to data stored in object/relational databases.

13. Short-term work (<12 months) should concentrate on developments for deploying cloud infrastructures and longer-term work should concentrate on their use.

14. The StratusLab distribution must be simple enough for users themselves to configure their own resources as a cloud.

15. The StratusLab distribution must allow both full-virtualization and para-virtualization to be used.

16. The cloud service must have a command line interface and a programmable API.

17. The cloud distribution must allow a broad range of grid and standard services to be run.

18. Quantitative performance evaluations must be done to understand the penalties in using virtualization.

19. The project must determine the criteria by which administrators and users can trust machine images.

20. The project should consider all features listed in the surveys as valid requirements.

21. Integration with site management tools is a critical short-term requirement.

22. The cloud implementation must scale to O(10000) virtual machines.

23. The implementation must sufficiently sandbox running machine images to prevent unintended or malicious behavior from affecting other machines/-tasks.

24. The project must create application benchmarks (CPU-Intensive, Simulation, Analysis, Filtering, Shared Memory, Parallel, and Workflow) to measure quantitatively the performance of the cloud implementation for realistic applications.

25. Performance benchmarks should also be created using packages like HEP-SPEC, Iozone, and iperf for CPU, disk IO, and network performance, respectively.

Glossary

Appliance	Virtual machine containing preconfigured software or services
DCI	Distributed Computing Infrastructure
EGEE	Enabling Grids for E-sciencE
EGI	European Grid Infrastructure
EGI-TF	EGI Technical Forum
GPFS	General Parallel File System by IBM
Hybrid Cloud	Cloud infrastructure that federates resources between organizations
iSGTW	International Science Grid This Week
NFS	Network File System
NGI	National Grid Initiative
Public Cloud	Cloud infrastructure accessible to people outside of the provider's organization
Private Cloud	Cloud infrastructure accessible only to the provider's users
VM	Virtual Machine
VO	Virtual Organization
VOBOX	Grid element that permits VO-specific service to run at a resource center
Worker Node	Grid node on which jobs are executed

References

[1] Amazon. Amazon Web Services. http://aws.amazon.com/.

[2] Apache Foundation. Hadoop. http://hadoop.apache.org/.

[3] Berkeley, University of California. BOINC. http://boinc.berkeley.edu/.

[4] CERN, FIO Group. HEP-SPEC: HEP benchmark for CPU performance. https://twiki.cern.ch/twiki/bin/view/FIOgroup/TsiBenchHEPSPEC.

[5] CernVM. CernVM. http://cernvm.cern.ch/cernvm/.

[6] Citrix Systems. Xen hypervisor. http://www.xen.org/.

[7] Eucalyptus Systems. Eucalyptus. http://www.eucalyptus.com/.

[8] gLite Consortium. gLite Middleware. http://glite.org/.

[9] HealthGrid. HealthGrid 2010. http://paris2010.healthgrid.org/.

[10] IBM. GPFS. http://www.ibm.com/systems/software/gpfs/.

[11] International Science Grid This Week. iSGTW. http://www.isgtw.org/?pid=1002626.

[12] IOzone. IOzone Filesystem Benchmark. http://www.iozone.org/.

[13] LCG France. LCG-FR. http://lcg.in2p3.fr/.

[14] Lustre. Lustre. http://wiki.lustre.org/.

[15] NLANR/DAST. Iperf. http://iperf.sourceforge.net/.

[16] OpenNebula Project. OpenNebula. http://opennebula.org/.

[17] Puppet Labs. Puppet. http://www.puppetlabs.com/.

[18] Quattor Community. Quattor Toolkit. http://quattor.org/.

[19] RedHat. KVM: Kernel Based Virtual Machine. http://www.linux-kvm.org/.

[20] SixSq Sàrl. SixSq website. http://www.sixsq.com/.

[21] StratusLab Project. StratusLab website. http://stratuslab.eu/.

[22] VENUS-C Project. VENUS-C. http://www.venus-c.eu/.

[23] VMware. VMware vSphere Hypervisor (ESXi)i. http://www.vmware.com/products/vsphere-hypervisor/index.html/.

[24] WLCG Project. WLCG. http://lcg.web.cern.ch/lcg/.

[25] Zoomerang. Zoomerang. http://www.zoomerang.com/.

A System Administrator Survey and Results

This appendix contains the questions from the system administrator survey along with the raw (or slightly processed) results. The results are shown in **bold**. When respondents were asked to rank something, the average result is shown. If "not used" was a possible answer, then the number of responses is also shown. Email addresses are not published here.

A.0 Welcome Page

The survey has **41 questions** on **8 pages** and takes **15 minutes** to complete. Your feedback is appreciated.

StratusLab will bring **cloud technology** to existing and new **grid resource centers**. The project focuses on the quality of the StratusLab cloud distribution and verifying the performance of real applications running on those clouds.

The survey determines the level of experience with virtualization and cloud technologies within the European e-infrastructures. It focuses on **grid resource center administrators** using or planning to use virtualization and/or cloud technologies. A companion survey focuses on end-users.

The participant's email address (optional) is the only personal information collected in the survey. You can opt-in to the StratusLab announcement list and indicate if we can contact you for follow-up of your responses. **Email addresses will not appear in any public summary of the results.**

A.1 Background Information (Page 1 of 8)

1. Subscribe (opt-in) to the StratusLab announcement list? This list is a low-volume list used to announce new versions of the StratusLab distribution, tutorials, etc.

   10: Yes

   12: No

2. May we contact you via email about your responses? (Default: No)

   11: Yes

   10: No

3. Email Address (if you responded yes to one of the questions above)
   Email addresses provided to dissemination activity.

4. What type of institute do you work for?

   12: Public Research Institute
   0: Private Research Institute
   6: Educational Institute
   2: Government Entity
   1: Large Enterprise
   1: Small or Medium Enterprise
   0: Not-for-profit organization

5. Country where your institute is located.

   9: France
   3: Germany
   3: Spain
   2: Switzerland
   1: Belgium
   1: Serbia
   1: Sweden
   1: United Kingdom
   1: United States

6. Specify your scientific or commercial domain of activity. (E.g. astronomy, bioinformatics, engineering, etc.)

   1: Bioinformatics
   0: Biology, Biomedical, Imaging, Complex Systems
   1: Chemistry, Material Science
   5: Comp. Science/Software Eng.
   0: Earth Sciences, Seismology, Meterology
   2: Education
   1: Finance
   12: Physics, High-Energy Physics

A.2 Resource Center Characteristics (Page 2 of 8)

7. Approximate number of physical hosts in your resource center.

   10: 16–100
   4: 101–500
   4: 501–1000
   1: 1001–5000
   3: 5001–20000

8. Indicate the primary operating system used on your physical hosts.

   19: RedHat-based Linux (RHEL, CentOS, Scientific Linux, ...)
   3: Debian-based Linux and derivatives (Debian, Ubuntu, ...)
   0: Other Linux (SUSE, Mandriva, ...)
   0: Unix (OpenSolaris, FreeBSD, ...)
   0: Windows
   0: MacOS X
   0: OS Independent (Java, Ruby, Python, ...)
   0: Other, please specify

9. What site configuration and management tools are you using at your resource center?

   6: None
   11: Quattor
   6: Puppet
   1: Cfengine
   0: LCFG
   1: Rocks
   1: Chef
   2: Other, please specify
   Home-grown system.

10. What file systems have you deployed?

   11: Local disk file systems
   17: Network File System (NFS)
   4: GPFS
   5: Lustre

5: Other, please specify
AFS (2), PNFS, HDFS, CFS

11. What grid middleware distribution(s) do you use at your site?

2: None
18: gLite
2: ARC
1: Unicore
5: Other, please specify
VDT, Globus Toolkit, Home grown, CiGri, BOINC

12. What grid services does your site provide?

3: None
19: Computing Element
9: DPM Storage Element
4: dCache Storage Element
2: SToRM Storage Element
6: File Transfer Service
12: LFC (File Catalog)
10: VOMS Server
12: VOBOX
2: AMGA Metadata Catalog
7: Other, please specify
Top Level BDII, User Interfaces, Quattor Server, Apel, UNICORE, Globus, xrootd, WMS/LB, Castor Storage Element

13. In which grid infrastructures does your site participate (EGI, an NGI, regional infrastructures, projects, etc.)?
Nearly all indicated EGI, either directly or through NGIs or closely related infrastructure projects. Others mentioned were Grid'5000 in France and the Nordic Data Grid Facility (NDGF).

A.3 Cloud/Virtualization Plans (Page 3 of 8)

14. Have you deployed or plan to deploy cloud or virtualization technologies? If not, why not?

15: Yes
7: No
Details: **Negative answers generally state that there is no demonstrated need for cloud technology yet. Others say that they already use virtualization is already in use for individual services but not as a "cloud".**

A.4 Cloud Experience and Plans (Page 4 of 8)

15. When are you planning to deploy virtualization/cloud technologies?

    6: Using Now
    1: Within 3 months
    2: Within 6 months
    6: Within 12 months
    0: Within 24 months
    0: After 24 months

16. If you provide a cloud or cloud-like service to users, provide a link to the service and/or public documentation.
    No useful responses.

17. Do you intend to use full (hardware) virtualization on your site?

    14: Yes
    1: No

18. Do you intend to use para-virtualization (software) on your site?

    10: Yes
    5: No

19. Which hypervisors do you use when deploying your virtual machines?

    10: KVM
    5: Xen (hardware virtualization)
    8: Xen (para-virtualization)
    8: VMWare ESXi
    0: VirtualBox
    0: OpenVZ
    2: Other, please specify
    HyperV

20. Rate how well the following virtual machine managers meet your requirements. (Scale: 1=Poor, 2, 3=Average, 4, 5=Outstanding, Not Used)

    3.6, 5: OpenNebula
    3.0, 4: Eucalyptus
    0: Nimbus
    2.5, 4: vSphere

4.0, 2: Xen Cloud Platform

21. If you have done a formal evaluation of any of the above virtual machine managers, provide links to the evaluation results.
    No useful responses.

22. Rate the importance of these types of cloud service interfaces to your users. (Scale: 1=Unneeded 2, 3=Useful, 4, 5=Mandatory)

    3.3: Programming API (Java, C++, Python, ...)

    3.5: HTTP(S)-based API (REST, XMLRPC, ...)

    3.0: Web Service API (SOAP, ...)

    3.9: Command Line Interface

    3.1: Browser-based Interface

23. Do you intend to run standard site services (e.g. web servers, mail servers, etc.) within a virtualized environment?

    10: Yes

    5: No

24. Do you intend to run grid services within a virtualized environment?

    13: Yes

    2: No

25. If there are grid or standard services you will continue to run on physical resources, which services and why?
    Most responses said that the performance of services will have to be evaluated on virtualized resources. Particular instances: Infiniband resources, non-dynamic services like Cluster File System, and databases.

A.5 Virtual Machines & Appliances (Page 5 of 8)

26. Who provides/creates the virtual machines or appliances that you use?

    13: Myself

    3: Known colleague in scientific/commercial community

    0: RightScale

    0: rPath

    4: CernVM

    0: Cloud Provider (e.g. stock Amazon images)

    2: Other, please specify
    "Official" group within community or project.

27. What tool(s) do you use to build customized virtual machines or appliances?

   4: None (don't build customized images)

   0: bitNami

   1: rBuilder

   0: Kameleon

   2: CernVM

   0: Cloud provider tools (e.g. Amazone build tools)

   9: Manual creation (e.g. from OS distribution media)

   3: Post-boot customization of stock images

   0: Other, please specify

28. Will you allow user-generated virtual machines to be deployed on your site? If so, what criteria do you use to determine if you can trust those virtual machines?

   5: Yes

   9: No

Details: **Responses cite issues of trust, but do not provide criteria.**

29. Indicate the primary operating system used used within your virtual machines.

   13: RedHat-based Linux (RHEL, CentOS, Scientific Linux, ...)

   2: Debian-based Linux and derivatives (Debian, Ubuntu, ...)

   0: Other Linux (SUSE, Mandriva, ...)

   0: Unix (OpenSolaris, FreeBSD, ...)

   0: Windows

   0: MacOS X

   0: OS Independent (Java, Ruby, Python, ...)

   0: Other, please specify

A.6 Cloud Features (Page 6 of 8)

30. Rate the importance of these general features. (Scale: 1=Not Important, 2, 3=Average, 4, 5=Essential)

   3.6: Near instantaneous start-up of virtual machines (responsiveness)

   3.5: Potential access to large pool of resources (elasticity)

   3.9: Guaranteed Quality of Service (QoS) for provided CPU

31. Rate the importance of these data-related features.

    3.9: File-based data access (storage and retrieval)
    4.0: Network-based data access (storage and retrieval)
    3.2: Creation and use of block storage devices (e.g. virtual disks)
    3.7: Guaranteed Quality of Service (QoS) for data storage

32. Rate the importance of these networking features.

    3.2: Allocation of public IP address
    3.4: Control over machine firewall (network accessibility)
    2.8: Automatic deployment of Virtual Private Network
    3.7: Guaranteed Quality of Service (QoS) for network access

33. Rate the importance of these features for virtual machine control.

    3.9: Customization of execution environment (e.g. customized virtual machine)
    3.9: Deployment of groups of virtual machines (e.g. entire batch system)
    3.4: Availability of standard virtual machines/appliances
    3.4: Ability to migrate virtual machines
    3.1: Ability to suspend virtual machines

34. Rate the importance of these monitoring features.

    4.2: Dashboard of running virtual machines
    4.2: Performance information for virtual machines
    4.1: Accessibility of complete accounting information
    3.7: Near real-time access to performance and/or accounting information

35. Rate the importance of cloud/virtualization integration with site management tools. **(Ave. = 4.7)**

    0: 1=Not important
    0: 2
    4: 3=Average
    4: 4
    7: 5=Essential

36. List important additional features for your use of a cloud or virtualized infrastructure.

    - **Management without the need of MS Windows machines**
    - **Scalability (>O(10000) VMs)**
    - **Automated network deployment and IP management**

A.7 Benchmarks (Page 7 of 8)

37. StratusLab will use the following application-level benchmarks to understand the performance of the distribution. Rate the importance of each benchmark. (Scale: Unimportant, 2, Average, 4, Critical)

    3.8: CPU-intensive: High-CPU requirements but little or no input and output data.

    3.7: Simulation: Small input, but significant output data with high-CPU requirements.

    3.9: Analysis: Large input data but relatively small output data.

    3.5: Filtering: Large input and output data.

    3.2: Shared Memory: OpenMP-like programs

    3.1: Parallel: MPI-like programs

    3.2: Workflow: Multiple interdependent tasks.

38. Describe any other application-level benchmarks StratusLab should consider.

    - **CPU performance (HEP-SPEC)**
    - **Disk performance (iozone)**
    - **Network performance (iperf)**

A.8 Use Cases and Encountered Problems (Page 8 of 8)

39. Describe any important use cases for the provision of a cloud or virtualized infrastructure.

    - **Standard batch worker nodes**
    - **no idea fro the moment**
    - **the usage of user/group provided images with customized SW and ability to inspect it**

40. Describe any major problems or issues you have encountered when using cloud or virtualized infrastructures.

    - **We were facing problems in relation to the integration to our local site monitoring tools. We use the "parent"-relation of nagios to create dependencies between a physical server and the VMs running on it. When a VM migrates from host A to host B, this change has to be forwarded to the nagios configuration. (that not really a major issue, but it's an issue ;-))**

- **Scalability of the provisioning system Scalability of the existing infrastructure (eg doubling the number of machines in a Computer Center is a non-trivial exercise) VM Performance VM Monitoring Integration into existing infrastructure**
- **- I/O Performance on block devices**
- **Security and accounting**
- **supported virtualization technology in base OS system**

41. Additional Comments
    No additional comments.

B User Survey and Results

This appendix contains the questions from the user survey along with the raw (or slightly processed) results. The results are shown in **bold**. When respondents were asked to rank something, the average result is shown. If "not used" was a possible answer, then the number of responses is also shown. Email addresses are not published here.

B.0 Welcome Page

The survey has **36 questions** on **8 pages** and takes **15 minutes** to complete. Your feedback is appreciated.

StratusLab will bring **cloud technology** to existing and new **grid resource centers**. The project focuses on the quality of the StratusLab cloud distribution and verifying the performance of real applications running on those clouds.

The survey determines the level of experience with virtualization and cloud technologies within the European scientific community. It focuses on **end-users of virtualization and cloud technologies**. A companion survey focuses on system administrators.

The participant's email address (**optional**) is the only personal information collected in the survey. You can **opt-in** to the StratusLab announcement list and indicate if we can contact you for follow-up of your responses. **Email addresses will not appear in any public summary of the results**.

B.1 Background Information (Page 1 of 8)

1. Subscribe (opt-in) to the StratusLab announcement list? This list is a low-volume list used to announce new versions of the StratusLab distribution, tutorials, etc.

   27: Yes

   28: No

2. May we contact you via email about your responses? (Default: No)

   28: Yes

   26: No

3. Email Address (if you responded yes to one of the questions above)
   Email addresses included in dissemination mailing lists.

4. What type of institute do you work for?

   31: Public Research Institute
   2: Private Research Institute
   14: Educational Institute
   3: Government Entity
   1: Large Enterprise
   3: Small or Medium Enterprise
   2: Not-for-profit organization

5. Country where your institute is located.

   26: France
   11: Greece
   5: Switzerland
   2: Netherlands
   2: United Kingdom
   2: United States
   1: Germany
   1: Ireland
   1: Italy
   1: Korea
   1: Norway
   1: Romania
   1: Spain
   1: Sweden

6. Choose the category that best describes you.

   23: End-User/Scientist: Takes advantage of virtualized/cloud infrastructure by using existing virtual machines (or "appliances") to run scientific or commercial applications.
   17: Developer/Engineer: People that port scientific applications to virtualized/cloud infrastructures and/or create customized virtual machines (or "appliances")
   3: Service Administrator: People that are responsible for running Virtual Organization (VO) services (as opposed to core grid or standard services) for a particular VO.

13: Computer Science Researcher: People investigating virtualization and cloud technologies themselves rather than looking to make use of those technologies for analyses in other fields.

7. Specify your scientific or commercial domain of activity. (E.g. astronomy, bioinformatics, engineering, etc.)

   15: Bioinformatics

   8: Biology, Biomedical, Imaging, Complex Systems

   3: Chemistry, Material Science

   16: Comp. Science/Software Eng.

   5: Earth Sciences, Seismology, Meterology

   0: Education

   0: Finance

   9: Physics, High-Energy Physics

B.2 Application Characteristics (Page 2 of 8)

8. Indicate the primary operating system used by your application.

   26: RedHat-based Linux (RHEL, CentOS, Scientific Linux, ...)

   13: Debian-based Linux and derivatives (Debian, Ubuntu, ...)

   3: Other Linux (SUSE, Mandriva, ...)

   1: Unix (OpenSolaris, FreeBSD, ...)

   6: Windows

   2: MacOS X

   4: OS Independent (Java, Ruby, Python, ...)

   1: Other, please specify **(Any Unix)**

9. Type(s) of executables(s) used in your application.

   36: Sequential

   26: Multi-threaded

   11: Shared Memory (e.g. OpenMP)

   21: Parallel (e.g. MPI)

10. Describe the overall control model for your multiple-job/task analyses.

   5: Only single jobs used for complete analysis

   5: Interactive control of running jobs/tasks

   14: Manual submission of independent jobs/tasks (batch)

11: Automatic submission of independent jobs/tasks (master/workers, parameter sweep)

16: Automatic orchestration of interdependent tasks (workflow)

5: Parallel execution of interdependent tasks (parallel)

11. Typical size of input data for a single job/task.

3: <10 kB

2: 100 kB

7: 1 MB

10: 10 MB

11: 100 MB

9: 1 GB

10: 10 GB

4: >100 GB

12. Typical size of output data for a single job/task.

3: <10 kB

3: 100 kB

5: 1 MB

11: 10 MB

15: 100 MB

8: 1 GB

6: 10 GB

5: >100 GB

13. Approximate time needed to run a single job/task on modern machine/CPU.

4: 1 Minute

10: 10 Minutes

14: 1 Hour

11: 10 Hours

6: 1 Day

11: >1 Day

14. Do you have privacy, contractual, or confidentiality constraints for data stored in the cloud? If yes, indicate the type of protection needed (access control, encryption, etc.)

21: Yes

35: No

Details: **Both access control lists and encryption are necessary. Access control was mentioned in nearly all responses; encryption was mentioned in about half.**

15. What are the most important mechanisms for data access in your application? (Leave blank if you don't know.)

24: POSIX file access

19: Remote File I/O

9: Relational database

6: Object database

4: Block storage

1: Other, please specify

B.3 Cloud/Virtualization Plans (Page 3 of 8)

16. Are you using or planning to use virtualization/cloud technologies? If not, why not?

43: Yes

13: No

Details: **In these responses, there were a couple that indicated that cloud technologies would be more interesting if people had access to those infrastructures and some knowledge of what could be accomplished. Other responses indicated that current facilities are sufficient for their needs.**

B.4 Cloud Experience and Plans (Page 4 of 8)

17. When are you planning to use virtualization/cloud technologies?

16: Using Now

6: Within 3 months

2: Within 6 months

9: Within 12 months

8: Within 24 months

2: After 24 months

18. If you are using or will use a cloud infrastructure, you will principally use:

12: Private (non-commercial) clouds only

4: Public (commercial) clouds only
12: Both private and public clouds
15: Your own resources configured as a cloud
0: None

19. How often do you use a cloud infrastructure?

18: Never
4: Once
6: Occasionally (<1 per mo.)
3: Monthly
4: Weekly
8: Daily

20. Rate the following public cloud service providers. (Scale: 1=Poor, 2, 3=Average, 4, 5=Outstanding, Not Used)

3.6, 16: Amazon Elastic Computing Cloud (EC2)
3.5, 11: Amazon Elastic Block Store (EBS)
3.7, 13: Amazon Simple Storage Service (S3)
3.4, 15: Google App Engine
2.6, 5: SalesForce
2.5, 4: ElasticHosts
2.6, 4: FlexiScale
2.7, 7: GoGrid
3.3, 7: Azure
3.2, 5: RightScale

21. If you use non-commercial clouds, provide the names and/or URLs for the services.
Several mentions of the Grid'5000 infrastructure.

22. Rate the importance of these types of cloud service interfaces. (Scale: 1=Unneeded, 2, 3=Useful, 4, 5=Mandatory)

4.2: Programming API (Java, C++, Python, ...)
3.4: HTTP(S)-based API (REST, XMLRPC, ...)
3.3: Web Service API (SOAP, ...)
4.1: Command Line Interface
3.6: Browser-based Interface

B.5 Virtual Machines & Appliances (Page 5 of 8)

23. Who provides/creates the virtual machines or appliances that you use?

    19: Myself

    19: Known colleague in scientific/commercial community

    0: RightScale

    1: rPath

    4: CernVM

    6: Cloud Provider (e.g. stock Amazon images)

    2: Other, please specify
    No image used.

24. What tool(s) do you use to build customized virtual machines or appliances?

    21: None (don't build customized images)

    0: bitNami

    0: rBuilder

    0: Kameleon

    3: CernVM

    9: Cloud provider tools (e.g. Amazon build tools)

    13: Manual creation (e.g. from OS distribution media)

    1: Other, please specify
    LYaTiss: http://www.lyatiss.com/

25. What criteria do you use to determine if you can trust a virtual machine or appliance created by someone else?
    No useful answers provided.

B.6 Cloud Features (Page 6 of 8)

26. Rate the importance of these general features. (Scale: 1=Not Important, 2, 3=Average, 4, 5=Essential)

    3.8: Near instantaneous start-up of virtual machines (responsiveness)

    4.2: Potential access to large pool of resources (elasticity)

    4.0: Guaranteed Quality of Service (QoS) for provided CPU

27. Rate the importance of these data-related features.

    4.2: File-based data access (storage and retrieval)

    4.1: Network-based data access (storage and retrieval)

3.2: Creation and use of block storage devices (e.g. virtual disks)

3.7: Guaranteed Quality of Service (QoS) for data storage

28. Rate the importance of these networking features.

3.5: Allocation of public IP address

3.7: Control over machine firewall (network accessibility)

3.5: Automatic deployment of Virtual Private Network

3.7: Guaranteed Quality of Service (QoS) for network access

29. Rate the importance of these features for virtual machine control.

4.2: Customization of execution environment (e.g. customized virtual machine)

4.0: Deployment of groups of virtual machines (e.g. entire batch system)

3.9: Availability of standard virtual machines/appliances

3.2: Ability to migrate virtual machines

3.3: Ability to suspend virtual machines

30. Rate the importance of these monitoring features.

3.7: Dashboard of running virtual machines

3.9: Performance information for virtual machines

3.8: Accessibility of complete accounting information

3.7: Near real-time access to performance and/or accounting information

31. List important additional features for your use of a cloud or virtualized infrastructure.

    - **Security and isolation of machines.**
    - **Near immediate availability of machine images.**

B.7 Benchmarks (Page 7 of 8)

32. StratusLab will use the following application-level benchmarks to understand the performance of the distribution. Rate the importance of each benchmark. (Scale: Unimportant, 2, Average, 4, Critical)

3.8: CPU-intensive: High-CPU requirements but little or no input and output data.

4.1: Simulation: Small input, but significant output data with high-CPU requirements.

3.8: Analysis: Large input data but relatively small output data.

3.8: Filtering: Large input and output data.

3.7: Shared Memory: OpenMP-like programs

3.8: Parallel: MPI-like programs

3.8: Workflow: Multiple interdependent tasks.

33. Describe any other application-level benchmarks StratusLab should consider.
No useful additional comments.

B.8 Use Cases and Encountered Problems (Page 8 of 8)

34. Describe any important use cases for your use of a cloud or virtualized infrastructure.

    - **Specialized services deployment (e.g. medical database). Secured/isolated processes (network virtualization and transparent encryption of all communications). Dynamic redeployment of cloud resources (along the execution of a workflow, group of nodes are needed for a limited period of time and can be replaced by different nodes/virtual images along time).**
    - **Thinks in parallel, executes sequentially and is always stuck in the bottleneck...**
    - **Clustering DNA sequencs**
    - **OLAP queries Data mining**
    - **To replace current any IT physical infrastructure**
    - **Long running statistical analysis (R engine) jobs. Cloud backed storage.**
    - **computing on demand for the scientific community**
    - **I would love to use a cloud within the European grid for portability reasons mostly, so we are able to easily integrate user models in native code without portability issue.**
    - **simulations and image processing**
    - **medical data**

35. Describe any major problems or issues you have encountered when using cloud or virtualized infrastructures.

    - **Time to deploy virtual nodes system images introduces a critical overhead.**
    - **Cannot access to UI servers or network errors.**

- **Small amount of memory per CPU**
- **No QoS with current solutions**
- **Runtime configuration when elastically scaling (e.g. adding nodes to a batch system) , also related to security.**
- **virtualisation cpu overhead (but minor), not gpu ”enabled”**
- **To run virtualized machines on EGEE we have to embed a statically compiled version of qemu. It provides full virtualization in a fairly slower manner than a paravirtualized system does. Providing kvm on every EGEE node would be a really nice feature for us.**
- **Quality of Service, intra-node network communication**

36. Additional Comments

    - **There is a clear separation between different aspects usually encompassed in the ”cloud” term: - virtualization: flexible dedicated images for dedicated application codes. - resources allocation: reservation of a fixed number of resources. Clouds usually deliver flexible resource allocations but very limited support, if any, to exploit these resources (scheduling, load balancing, cost estimation...) which is critical from a user point of view.**
    - **Good luck with your survey! Bring us beautiful parallel systems :D**
    - **Network QoS is central**
    - **gpgpu techs are becoming popular in scicomp/imaging.**
    - **don’t really know this technology, sorry ...**